Uma Abordagem de Extração de Terminologia para a Construção de uma Representação Atributo-valor a Partir de Documentos Não Estruturados*

Daniel de Faveri Honorato[1], Maria Carolina Monard[1], and Huei Diana Lee[2] and Wu Feng Chung[2]

[1]Departamento de Ciência da Computação
Instituto de Ciências Matemáticas e de Computação (ICMC) da Universidade de São Paulo (USP)
São Carlos - São Paulo - Brasil CEP 13560-970
[2]Laboratório de Bioinformática
Universidade Estadual do Oeste do Paraná (UNIOESTE)
Parque Tecnológico Itaipu (PTI)
Foz do Iguaçu - Paraná - Brasil CEP 85856-970
{dfh,mcmonard}@icmc.usp.br

Resumo Neste trabalho é feita uma proposta metodológica, a qual pode ser aplicada tanto automaticamente como semi-automaticamente com ajuda de um especialista do domínio, que utiliza uma abordagem híbrida de extração de terminologia para realizar o mapeamento de documentos não estruturados para uma tabela atributo-valor. Essa abordagem auxilia na identificação de relacionamentos entre as entidades, e permite filtrar palavras e frases que ocorrem acima de um limiar por meio da aplicação de medidas estatísticas. A metodologia foi implementada em um ambiente computacional e foi avaliada na estruturação automática de uma coleção de 6000 documentos de Endoscopia Digestiva Alta descritos em língua natural. Os resultados obtidos mostram que ela é adequada para reduzir o tempo de atuação do especialista na análise de grandes quantidades de documentos não estruturados.

Key words: Text pre-processing, text mining, terminology extraction

1 Introdução

O desenvolvimento e a utilização de tecnologias para a aquisição e o armazenamento de dados, nas mais diversas áreas do conhecimento, têm permitido o acúmulo de dados em uma velocidade maior que a capacidade humana possui para processá-los. Esses dados podem estar representados em diferentes formatos, sendo que um dos formatos bastante utilizados é o formato textual não estruturado. Para que esses dados textuais brutos possam tornar-se úteis, é

* Trabalho desenvolvido com o apoio da Fundação Parque Tecnológico Itaipu e CNPq.

necessário que eles sejam representados de maneira apropriada para a extração de padrões, tal que um modelo que represente o conhecimento embutido nesses dados possa ser construído. Uma das maneiras de alcançar esse objetivo é por meio do processo de Mineração de Textos — MT [2]. Esse processo consiste, basicamente, em três fases principais: (1) pré-processamento dos textos; (2) extração de padrões; e (3) pós-processamento. Na fase de pré-processamento, o conjunto de textos é transformado para uma representação adequada para ser utilizada pelos algoritmos de extração de padrões. Geralmente, os documentos são representados em uma tabela atributo-valor, na qual palavras selecionadas do conjunto de documentos são transformadas em atributos. Essa tabela é utilizada na fase de extração de padrões, os quais são analisados e validados na fase de pós-processamento.

A proposta apresentada neste trabalho faz parte de um projeto de pesquisa mais amplo, desenvolvido conjuntamente por pesquisadores do Laboratório de Bioinformática — LABI[1] — da Universidade Estadual do Oeste do Paraná e do Laboratório de Inteligência Computacional — LABIC[2] — do ICMC/USP. No LABI foi proposta uma metodologia [4] para transformação de informações descritas em laudos médicos não estruturados em tabela atributo-valor, baseada nas seguintes duas propriedades das informações encontradas em diversos tipos de documentos textuais, entre eles, laudos médicos: 1 - as informações são descritas utilizando um vocabulário controlado; e 2 - as informações consistem de frases assertivas simples. Entretando, essa metodologia requer uma intensa interação com o especialista do domínio, o que não é sempre possível. Esse problema motivou o desenvolvimento da metodologia proposta neste trabalho.

O objetivo deste trabalho consiste da proposta, do desenvolvimento e da implementação de uma metodologia geral para o pré-processamento de documentos não estruturados, ou seções específicas desses documentos, que tem como resultado a representação em uma tabela atributo-valor das informações contidas nesses documentos. Um dos principais objetivos dessa metodologia é diminuir, sempre que possível, a intervenção do especialista, fornecendo, nas diversas fases da metodologia, informações que facilitam o trabalho a ser realizado pelo especialista, que é o responsável pela inclusão do aspecto semântico das informações. De fato, a metodologia pode ser aplicada, de modo automático, sem intervenção do especialista, considerando somente o aspecto morfo-sintático das informações, *i.e.* sem levar em conta o aspecto semântico. Como produto deste trabalho foi desenvolvido o ambiente computacional *Term Pattern Discover* — TP-Discover — o qual implementa todas as fases da metodologia desenvolvida [5], [6].

Na literatura são encontrados poucos trabalhos de aplicação de métodos da área de mineração de textos sobre laudos médicos. No trabalho de [3] são descritos alguns desses trabalhos.

[1] http://labi.pti.org.br/
[2] http://labic.icmc.usp.br/

Este trabalho está organizado da seguinte maneira: na Seção 2 é apresentada a metodologia desenvolvida; na Seção 3 a avaliação experimental realizada e, finalmente, na Seção 4 são apresentadas as conclusões.

2 Metodologia

A metodologia proposta, a qual foi implementada no ambiente computacional TP-Discover, é composta por cinco fases: (1) Pré-processamento; (2) Extração de terminologia; (3) Identificação de atributos; (4) Construção do dicionário; e (5) Construção da tabela atributo-valor. Essas fases, descritas a seguir, estão ilustradas na Figura 1.

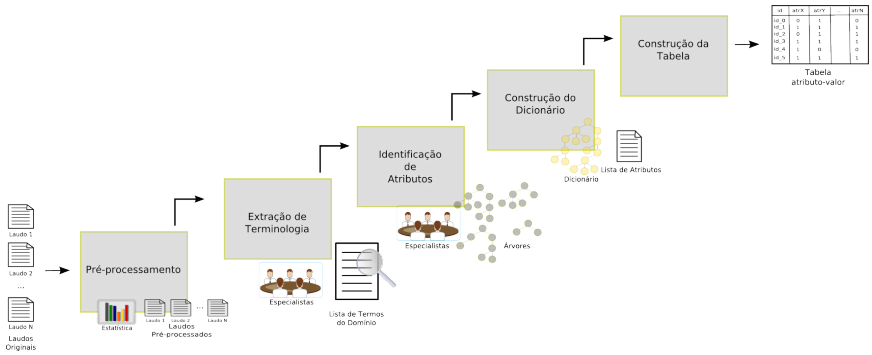

Figura 1. Metodologia desenvolvida

Pré-processamento – A fase de pré-processamento consiste de sete etapas, as quais são aplicadas sobre o conjunto de laudos do domínio que está sendo trabalhado. Essas etapas correspondem à preparação do *corpus* por meio de tarefas como: divisão do documento de acordo com as seções que ele contém; construção do Conjunto de Frases Únicas — CFU — que consiste de todas as frases diferentes do conjunto de documentos; remoção de *stopwords* que são palavras consideradas não relevantes para a análise do texto; transformação para minúsculo; correção ortográfica; aplicação de substituições tais como sinônimos ou frases que mapeiam mais de um evento; e aplicação do lematizador, o qual transforma verbos para a forma infinitiva e substantivos e adjetivos para masculino singular.

Extração de Terminologia – A terminologia é uma área de conhecimento e de práticas relacionados a termos técnico-científicos [1]. Dois exemplos de terminologia seriam a terminologia da medicina e a terminologia usada pelos especialistas da computação. Geralmente, sistemas de Extração de Terminologia — ET — utilizam conhecimento estatístico, lingüístico ou híbrido [7], [8]. Na abordagem estatística, termos são reconhecidos a partir de suas freqüências de ocorrências em um *corpus*. Na abordagem lingüística, o *corpus* é etiquetado com um etiquetador do tipo *part-of-speech* e termos com determinadas classes gramaticais são filtrados por meio de casamento de padrões. A abordagem híbrida combina técnicas da abordagem lingüística e estatística.

Neste trabalho, a fase de ET tem por objetivo extrair termos do domínio que são freqüentemente utilizados pelo profissional no mapeamento de informações.

Por exemplo, suponha que o profissional sempre mapeie uma informação no documento, cuja seqüência de palavras inicia-se com a palavra X. Portanto, no conjunto de documentos pode existir, no contexto da palavra X, diversas seqüências, tais como X A B, X B e X C D, e o objetivo é determinar as mais apropriadas para serem consideradas como unidades terminológicas. Neste trabalho, para a identificação das unidades terminológicas é adotada uma abordagem híbrida seguida da aplicação de algumas heurísticas sobre o conjunto de termos identificados. Um fator primordial para a escolha do método híbrido está no fato de que os termos de interesse, por exemplo a palavra X, na maioria dos casos pertencem à uma determinada classe gramatical e, sobre esses termos, é então calculada a freqüência de cada um. Neste trabalho, após uma análise morfo-sintática prévia dos termos do domínio, foi decidido gerar são geradas duas listas, uma lista de unigramas contendo palavras que casam com a classe gramatical N (*substantivo*) e outra lista de bigramas contendo palavras que casam com palavras da classe gramatical N N.

Um dos problemas identificados nas listas resultantes da aplicação do método híbrido, é que muitos termos possuem baixa freqüência ou muitos termos que aparecem em uma determinada lista de unigramas também fazem parte de algum bigrama da lista de bigramas. Assim, com o objetivo de encontrar unidades terminológicas mais apropriadas do domínio, neste trabalho foram propostas algumas heurísticas que auxiliam na redução do número de termos identificados. Essas heurísticas trabalham sobre uma lista qualquer de unigrama e uma lista qualquer de bigramas e levam em consideração a freqüência dos termos presentes nas duas listas. Nessas heurísticas é utilizado o parâmetro Alpha, o qual, a partir da lista de unigramas e da lista de bigramas, permite favorecer a escolha de um unigrama ou de um bigrama para fazer parte da lista de termos candidatos final. Depois de aplicadas as heurísticas, são removidos da lista de termos candidatos todos os termos (unigramas ou bigramas) que possuem freqüência menor ou igual a um limiar Theta, definido pelo usuário, em relação ao número de documentos. A lista de termos candidatos final é utilizada na próxima fase da metodologia.

Identificação de Atributos, Construção do Dicionário e da Tabela – A identificação dos atributos é realizada em três etapas: definição dos termos que serão utilizados como raíz das árvores; geração das árvores; e identificação dos atributos a partir das árvores geradas. A identificação de termos raíz pode ser realizada de duas maneiras: automática ou manual. No modo automático, todos os termos identificados na fase de ET são considerados termos raíz. No modo manual, uma análise, junto com o especialista, pode ser realizada sobre a lista de termos com o intuíto de identificar os termos que realmente serão utilizados no mapeamento de informações, descartando termos que não são de interesse. Depois de definidos os termos raíz das árvores, é executado o algoritmo de geração de árvores, lembrando que para cada termo considerado como unidade terminológica é gerada uma árvore cuja raíz é definida por esse termo. As árvores geradas possuem uma estrutura semelhante à árvore ilustrada na Figura 2. Nessa árvore o termo α, onde o tamanho de α é dado por $1 \leq |\alpha| \leq 2$, corresponde ao nó raíz identificado pelo método de extração de terminologia, os termos β e γ

são filhos de α e δ é filho de β. Na árvore, todos os filhos possuem um número de palavras maior ou igual a 1 e mapeiam as palavras que aparecem no contexto de um termo raíz. Por exemplo, se a palavra `coloração` e `esbranquiçada` sempre aparecem juntas, elas serão colocadas em um mesmo nó filho, caso contrário serão colocadas em nós diferentes. Na árvore gerada também são armazenadas as freqüências em que dois termos ocorrem juntos. Por exemplo, a freqüência com que α e β aparecem juntos é $f_{\alpha\beta}$, já a freqüência com que $\alpha\beta\delta$ aparecem juntos é $f_{\alpha\beta\delta}$, sendo $f_{\alpha\beta\delta} < f_{\alpha\beta}$. Essa relação entre as freqüências se verifica em todos os ramos da árvore.

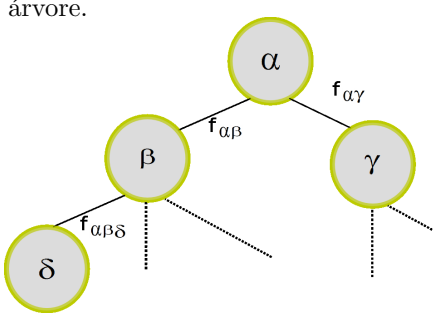

Figura 2. Árvore gerada

Após a construção do conjunto de árvores, é realizada a identificação dos atributos que irão compor a tabela atributo-valor. A identificação de atributos pode ser realizada com o auxílio do especialista (manual) ou automaticamente, na qual todos os ramos que possuirem freqüência maior ou igual a um limiar são considerados atributos. Por exemplo, considere o limiar l e a árvore da Figura 2. Na identificação automática de atributos, primeiramente é verificado se $f_{\alpha\beta} \geq l$ e, se for, é gerado o atributo $\alpha\beta$. Depois é verificado se $f_{\alpha\beta\delta} \geq l$ e, se for, é definido $\alpha\beta\delta$ como atributo. Por último é verificado se $f_{\alpha\gamma} \geq l$ e, se for, é definido $\alpha\gamma$ como atributo.

Depois de identificados os atributos, eles são inseridos em um dicionário, denominado de dicionário de conhecimento, o qual é representado utilizando uma estrutura de dados *trie* para facilitar a busca. Finalmente, o processo de preenchimento da tabela é realizado por meio de ciclos de pesquisa entre as informações dos documentos e os atributos presentes no dicionário de conhecimento construído. Nesse processo, se a seqüência de termos definida por um determinado atributo do dicionário for identificada em um documento, o valor desse atributo é preenchido com 1 (presente). Se a seqüência não for identificada no documento, o atributo é preenchido com 0 (ausente). Esse processo é repetido para todos os documentos do conjunto.

3 Avaliação Experimental

Foram realizados vários experimentos considerando o aspecto quantitativo, nos quais a metodologia foi avaliada utilizando diferentes valores dos parâmetros. Também foi realizada uma avaliação de qualidade da lista de termos identificados pelo método híbrido de extração de terminologia desenvolvido. A metodolo-

gia foi aplicada a um conjunto de laudos médicos composto por 6000 laudos de Endoscopia Digestiva Alta — EDA — obtidos no Hospital Municipal de Paulínia do estado de São Paulo. Os laudos são organizados em cinco seções, contendo informações do segmento de esôfago, estômago e duodeno, assim como informações referentes à realização de biópsia e conclusões do exame. Nas três primeiras seções e na última as informações estão descritas em língua natural. As informações sobre biópsia estão estrututras.

Para realizar os experimentos, foram analisadas duas seções dos laudos de EDA, especificamente as informações do esôfago e estômago. Em ambos os casos foram obtidos resultados semelhantes. Por falta de espaço, são apresentados somente os resultados da seção do esôfago. Também, foi decidido avaliar os resultados da aplicação da metodologia sem a intervenção do especialista, ou seja, considerando apenas o aspecto morfo-sintático. Assim, sem dúvida alguma, os resultados experimentais referem-se a resultados obtidos no *pior caso*, pois o aspecto semântico não foi considerado. Como os resultados necessitam ser avaliados não somente quantitativamente mas também qualitativamente, o conhecimento do especialista somente foi utilizado para calcular os valores de precisão e *recall* na fase de extração de terminologia, para os quais é necessário conhecer os termos de interesse do domínio.

Para avaliar a metodologia proposta no pior caso, foram realizados vários experimentos utilizando quatro diferentes conjuntos de documentos relacionados ao esôfago, os quais foram amostrados com reposição do conjunto total de 6000 documentos disponíveis. Na Tabela 3 é mostrado o número de documentos no conjunto de treinamento (Tr) e teste (Te) de cada um desses quatro experimentos. No Exp1 foram amostrados 1500 documentos do conjunto to-

Tabela 1. Configuração dos experimentos realizados

Id. Experimento	Tr	Te
Exp1	500	1000
Exp2	1000	1000
Exp3	2000	1000
Exp4	4000	1000

tal de documentos, os quais foram divididos em 500 para treinamento e 1000 para teste. No Exp2 foram amostrados 2000 do conjunto total de documentos e foram divididos em 1000 para treinamento e 1000 para teste, e assim por diante, utilizando sempre conjuntos diferentes de documentos em cada experimento.

Na fase de extração de terminologia foram utilizados diferentes valores para os parâmetros Alpha e Theta a fim de observar a influência deles na identificação dos termos de interesse do conjunto de laudos. Para cada experimento identificado na Tabela 3, foram utilizadas nove variações dos valores de Alpha e Theta. Para o parâmetro Alpha foram utilizadas três variações, 100%, 95% e 90%. Para cada Alpha foram utilizadas três variações de Theta, 5%, 10% e 20%, ou seja, no primeiro experimento foram utilizados os valores de Alpha=100% e Theta=5%. No segundo foram utilizados Alpha=100% e Theta=10% e assim sucessivamente. Os experimentos foram realizados de modo a extrair as seguintes informações: número de termos raiz; número de atributos gerados utilizando limiares de poda da árvore iguais a 0%, 5%, 10% e 20%; e taxas de preenchimento da tabela

atributo-valor gerada a partir do conjunto de treinamento e conjunto de teste, respectivamente.

A primeira tarefa antes de gerar as árvores foi realizar a identificação de unidades terminológicas na fase de extração de terminologia. A partir do conjunto de documentos etiquetados foram extraídos os termos pertencentes à classe gramatical N e N N. Sobre a lista de termos extraídos foram aplicadas as heurísticas que usam os valores de Alpha e Theta correspondentes. Após realizar a identificação desses termos (unidades terminológicas), os mesmos foram utilizados como termos raiz para gerar as árvores a partir das quais são identificados os atributos. Depois de geradas as árvores, foi realizada a identificação de atributos que compõem a tabela atributo-valor. Nos experimentos foram utilizados os limiares de poda das árvores 0%, 5%, 10% e 20%, para os quais utilizar o limiar 0% corresponde a transformar todos os ramos das árvores geradas em nomes de atributos. Na Figura 3 é ilustrado o gráfico que mostra a relação entre o número de termos raiz identificados e o número de atributos gerados utilizando o limiar de poda 0%, para os experimentos realizados. Nesse gráfico, no eixo x

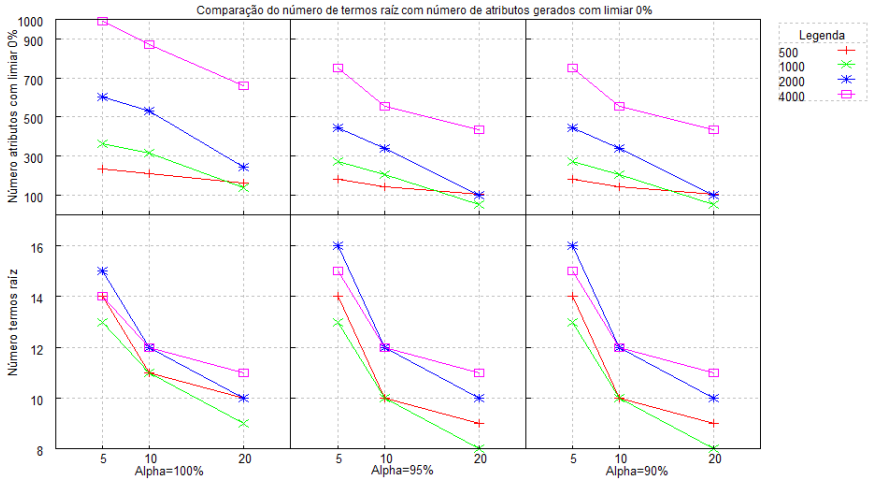

Figura 3. Relação entre número de termos raiz e número de atributos gerados utilizando limiar igual a 0%

estão representados os resultados obtidos com os valores de Alpha 100%, 95% e 90% e, para cada valor de Alpha estão representadas as variações de Theta (5%, 10%, 20%). É possível observar que usando as variações Alpha = 95% e Alpha = 90%, os resultados foram iguais, *i.e.*, foram identificados os mesmos termos. Isso significa que no restante dos experimentos da seção de esôfago, os resultados correspondentes a esses valores de Alpha são todos iguais. É possível observar no gráfico que quanto maior o número de documentos do conjunto de treinamento, maior o número de atributos identificados. Essa característica deve-se ao fato de que quanto mais documentos, mais variações existirão no contexto de um determinado termo raiz e, consequentemente, mais ramos serão criados a partir desse termo.

Para avaliar o número de atributos que são considerados em relação ao número total identificado com o limiar de poda igual a 0%, foram utilizados os limiares de 5%, 10% e 20%. Por falta de espaço, nas Figuras 4 e 5 são ilustrados somente os gráficos de 5% e 20%, respectivamente, que mostram a proporção com relação ao limiar 0%, de atributos considerados usando esses limiares, para os valores de Alpha 100%, 95% e 90%. Esses gráficos referem-se a proporção de atributos e não ao número de atributos, o qual é dado pelo produto do número total de atributos gerados utilizando o limiar 0% — Figura 3 — multiplicado pela taxa correspondente — Figura 4.

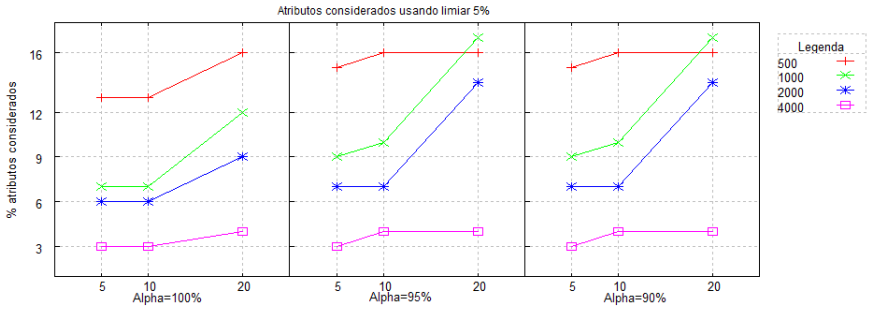

Figura 4. Taxa de atributos considerados utilizando limiar igual a 5%

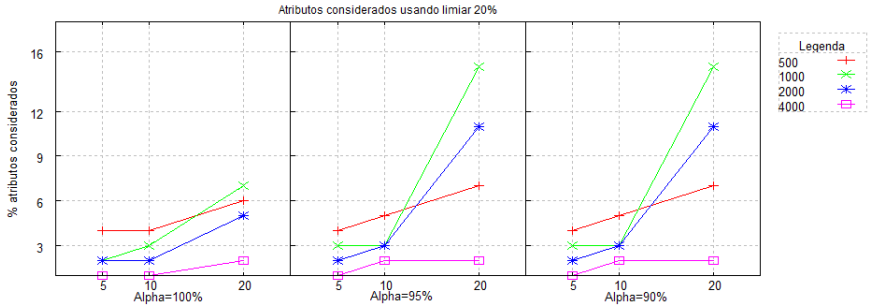

Figura 5. Taxa de atributos considerados utilizando limiar igual a 20%

No gráfico no qual foi utilizado o limiar de 5%, a taxa de atributos considerados em relação ao número total gerado é maior do que usando o limiar de 20%, uma vez que grande parte dos ramos das árvores possuem freqüência maior ou igual que 5%. É importante observar que para o conjunto de treinamento de 4000 laudos, a taxa de atributos considerados é menor pois, conforme observado na Figura 3, para esse experimento foi gerado um número muito maior de atributos, o que indica que foram geradas árvores com muitos ramos de freqüência menor que 5%. Por outro lado, a taxa de atributos considerados para o conjunto de treinamento de 500 laudos é maior, uma vez que o número de atributos gerado usando o limiar 0% é menor. Embora a taxa de atributos considerados de um experimento para outro seja grande, a variação do número de atributos não é grande. Conforme mencionado, isso se deve ao fato de que, com um conjunto

maior de documentos para treinamento, são geradas árvores com muitos ramos com freqüência baixa, enquanto os ramos com freqüência maiores estão próximas à raiz da árvore. Quando se aumenta o limiar de poda, somente serão considerados atributos próximos da raiz. Assim, o número total de atributos considerados utilizando um limiar de 20% é menor que o limiar 10% que é ainda menor que com um limiar de 5%.

Após identificados os atributos utilizando os diferentes limiares (0%, 5%, 10% e 20%) foi realizada a construção do dicionário e o preenchimento da tabela atributo-valor. Os resultados detalhados encontram-se em [3]. Como esperado, usando limiar 0%, a taxa de preenchimento é bastante baixa, ou seja, tem-se uma tabela bastante esparsa. Por outro lado, conforme o valor do limiar aumenta (5%, 10% e 20%), a taxa de preenchimento é incrementada, uma vez que o número de atributos é menor para limiares maiores, ou seja, estão sendo considerados como atributos apenas seqüências de palavras que têm freqüências maiores. Para todos os casos, para o limiar 20%, a taxa de preenchimento se estabiliza em aproximadamente 90%.

Nesses experimentos, também foram avaliados os valores de precisão e *recall* da lista de unidades terminológicas identificadas na fase de extração de terminologia. Para isso, foi utilizada uma lista de referência considerando apenas informações contidas na seção do esôfago dos documentos, a qual foi fornecida pelo especialista. Conforme foi observado, nos experimentos não ocorreram variações muito grandes nas medidas de precisão e *recall* entre os diferentes e correspondentes experimentos. Porém, nota-se que o melhor *recall* (55%) foi alcançado utilizando o valor de Alpha igual a 100% nos experimentos Exp1 e Exp4, e usando Theta igual a 5% e 10%, e a melhor precisão (60%) foi alcançada utilizando Alpha igual a 100% e Theta igual a 20% no Exp1. Deve ser levado em conta que devido ao uso de um vocabulário controlado, o uso de conjuntos de laudos de tamanhos diferentes para a extração de terminologia de interesse não influenciou muito e, em alguns casos, até degradou a precisão, pois termos não interessantes foram colocados na lista devido a sua alta freqüência no conjunto de documentos. Analisando a lista de referência do conjunto de documentos correspondentes a seção de esôfago, foi possível observar que a maioria dos termos fornecidos são unigramas e, usando Alpha igual a 100%, a identificação desses termos foi favorecida.

4 Conclusões

Neste trabalho foi apresentada uma metodologia para a transformação de informações não estruturadas, que foram registradas utilizando um vocabulário controlado e formadas por frases assertivas simples, para uma representação atributo-valor. Essa metodologia pode ser aplicada considerando apenas o aspecto morfo-sintático das informações ou pode utilizar conhecimento semântico, caso o especialista participe do processo. A identificação dos termos que serão utilizados como raiz das árvores pode ser controlada pelo usuário utilizando dois parâmetros, Alpha e Theta, os quais, na implementação realizada, permitem privilegiar termos unigramas ou bigramas. Após construídas as árvores, o

usuário pode definir um limiar de poda para controlar o número de atributos gerados. Um aspecto importante da metodologia é que ela não necessita de recursos externos de conhecimento do domínio, tais como dicionários de termos, regras semânticas e ontologias do domínio. Para ser aplicada, é utilizada somente a informação contida nos documentos. A metodologia foi avaliada no *pior caso*, ou seja, considerando apenas o aspecto morfo-sintático das informações, mostrando bons resultados. Caso o especialista participe do processo, o aspecto semântico pode ser considerado e os resultados serão certamente superiores. Uma outra contribuição importante deste trabalho é o projeto e implementação do ambiente computacional TP-Discover, no qual está implementada a metodologia proposta. Esse ambiente permite, por meio de interfaces amigáveis, configurar os diversos parâmetros que são utilizados na aplicação da metodologia.

Agradecimentos. Os autores agradecem Antonio Pietrobom Neto, do Hospital Municipal de Paulínia, por disponibilizar o conjunto de laudos utilizado nos experimentos.

Referências

1. da Graça Krieger, M. (2005). Terminologias em construção: procedimentos metodológicos. In *VIIII Congresso Internacional da ABECAN*, Rio Grande do Sul, Brasil.
2. Feldman, R. and Sanger, J. (2006). *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press, Nova Iorque, EUA.
3. Honorato, D. D. F. (2008). Metodologia de transformação de laudos médicos não estruturados e estruturados em uma representação atributo-valor. Dissertação de Mestrado, ICMC-USP. `http://www.teses.usp.br/teses/disponiveis/55/55134/tde-10062008-154826/publico/dissertacaoDanielHonorato.pdf`
4. Honorato, D. D. F., Cherman, E. A., Lee, H. D., Monard, M. C., and Wu, F. C. (2007). Construção de uma representação atributo-valor para extração de conhecimento a partir de informações semi-estruturadas de laudos médicos. In *Anais da XXXIII Conferencia Latinoamericana de Informatica*, pages 1–12, San José, Costa Rica.
5. Honorato, D. D. F. and Monard, M. C. (2008a). Descrição de uma metodologia de mapeamento de informações não estruturadas em uma representação atributo-valor. Technical Report 317, ICMC-USP. `http://www.icmc.usp.br/~biblio/BIBLIOTECA/rel_tec/RT_317.pdf`.
6. Honorato, D. D. F. and Monard, M. C. (2008b). Descrição do ambiente computacional tp-Discover para mapear informações não estruturadas em uma tabela atributo-valor. Technical Report 318, ICMC-USP. `http://www.icmc.usp.br/~biblio/BIBLIOTECA/rel_tec/RT_318.pdf`.
7. Pantel, P. and Lin, D. (2001). A statistical corpus-based term extractor. In *Proceedings of the 14th Biennial Conference of the Canadian Society on Computational Studies of Intelligence: Advances in Artificial Intelligence*, pages 36–46, Londres, Reino Unido.
8. Paulo, J. L., Correia, M., Mamede, N. J., and Hagège, C. (2002). Using morphological, syntactical, and statistical information for automatic term acquisition. In *PorTAL '02: Proceedings of the Third International Conference on Advances in Natural Language Processing*, pages 219–228, Londres, Reino Unido.